CAAP Quarterly Report (Oct-Dec 2019)

Date of Report:	January 31, 2020
Prepared for:	U.S. DOT Pipeline and Hazardous Materials Safety Administration
Contract Number:	Cooperative Agreement #693JK31850011CAAP
Project Title:	Development of a prediction model for pipeline failure probability based on learning from past incidents and pipeline specific data using artificial neural network (ANN)
Prepared by:	Texas A&M Engineering Experiment Station
Contact Information:	Joshua Arnold
For quarterly period ending:	December 31, 2019

Business and Activity Section

(a) Contract Activity

The contract activities have been continued.

(b) Status Update of Past Quarter Activities

Dr. Noor Quddus (Assistant Research Engineer) has been leading the team for the project. Graduate students working in the project are Guanyang Liu (3rd year PhD student) and Pallavi Kumari (3rd year Ph.D. student). An undergraduate researcher, Mason Boyd, has been working since last Summer.

(c) Cost share activity

Due to some unforeseen difficulties, financial information is not available at this moment. In the next quarter, all cost sharing activity will be updated.

(d) Task 1: *Development of methodology for creating root cause analysis reports*
Task 2: Selection of training samples and development of the learning algorithm

Both Task-1 and Task-2 have been continued.
Detailed discussion and descriptions for the following:

1. Background and Objectives in the first year

Pipeline failure incidents are reported by incident reporting system and some incidents are investigated regulated under pipeline safety regulations. Pipeline incident reports typically collects apparent causes that limits itself to direct causal factors without providing sufficient details of underlying cause of the incident. Different root cause analyses are usually employed to identify root causes and investigation reports express these causes in a variety of ways though certain management/organization related causes may be common in several incidents. It makes the development of any predictive model difficult, since there will be no uniformity of the input data to the predictive model. So, it would be essential to determine a reference that defines what will actually be termed as root cause. Using taxonomy so that similar terms are used to refer to related root causes can help tackle this issue. Then along with other data that contributed to past incidents will be used to develop the artificial neural network (ANN) model as a predictive model.

The overall objective is to develop a knowledge based predictive model to assess pipeline failure through:

a. Learning about causes behind pipeline failure: Conducting root cause analysis of past incidents to identify those factors that have to potential to contribute to failure. The findings are to be specific to the extent that they can be applied into a predictive model.
b. Implementation of learning to predict failure: Utilizing the learnings about contributing factors behind pipeline failure to develop a predictive model based on artificial neural networks that monitors current existing conditions to determine dynamic failure probability of a pipeline

The first challenge would be to build a set of cue words or taxonomy so that root causes analysis conducted for different incidents identify similar causes using similar terms and these causes will have to be identified in terms of measurable deviations/indicators so that they can later on be compared with deviations existing in a system to understand if the system is reaching an unsafe state. If a set of cue words are developed to produce all reports, extraction of information using automated systems based on text mining or data mining can be used. Task 1 will focus on determining how root cause analysis can be reported so that the information can be applied for prediction based on current condition.

ANN offers great potential for the development of a monitoring system based on past records while overcoming the limitations of the past attempts. An ANN model for the prediction of failure of pipelines based on findings from pipeline incident records, incident investigation report, and other available resources. The suitability of ANN for this purpose lies in its ability to do the following: learn from past records to produce a predictive model, model complex non-linear behavior that may exist in any socio-technical system, recognize or classify patterns in behavior and interaction of various contributing factors, and tolerate noises and deal with large data. They are particularly useful when there is no prior knowledge about how the variables interact since ANN models develop an understanding of the relations based on information provided during training. Thus, past findings can be utilized to train the ANN to recognize the relations between variables.

2. Analysis in the Quarters

In the current study, pipeline incidents have been analyzed. The study focuses to analyzing incident data and investigation reports for root cause identification and development of an artificial neural network model. The study started with background data analysis of the PHMSA incident data collection methodology and compared that with other pipeline incident databases. It considers the incident descriptions as reported in the incident reports for more insight. A natural language processing (NLP), a sub-field of artificial intelligence, model is being developed to analyze the incident descriptions. The model is also being tested on the incident investigation reports. The objective is to identify the root cause described in the very long text, which is humanly impossible to extract. On separate effort, ANN model has been developed and incident data have been analyzed to identify the suitable input data for the ANN model.

Data gathering: A small dataset from the enormous amount of data that PHMSA collected over the years has been selected for the current study. Incident data of hazardous liquid (HL) and natural gas transmission and gathering (GTG) have been gathered as they are the source of a lot of major incidents with higher incident rates. After 2010 incident reporting system has been updated and this dataset has vast amount of useful data. To avoid inconsistency in the input dataset, only incident data for 2010-2019 are being used for the current study.

Literature review: Comprehensive literature review of pipeline incidents has been conducted. A spread-sheet containing information of 39 articles and a summary of a few them have been shared. However, there are a lot more articles available in the literature and more literature review will conducted targeting specific objectives in due time.

Comparison of causal factors with other datasets: The incident data from other pipeline incident data sources e.g., Canada National Energy Board (NEB) and European Gas Pipeline Incident Data Group (EGIG) are also gathered. The causal factor categories reported in the PHMSA incident reports have been compared with that of other pipeline incident datasets. The preliminary assessment provided an overall status of global pipeline industry and provided guidance on how valuable practices from other incident sources can be utilized in the current study.

Pipeline failure data analysis: To understand the factors that contributed to pipeline failure, failure data were analyzed and compared from different perspective. Relationships among causal factor as reported in the incident reports, background factors (factors that have association with pipeline incident, but do not contribute directly to the incident; such as pipeline diameter, commodity transported), and root factors (underlying factor of the incident; management of organizational causes) have been explored. They provide us insight about what information to extract from the incident investigation or incident description and forms the base line of ANN input variables.

Development of natural language processing (NLP) model: The incident descriptions (3,616 incident counts) as reported in the incident report (HL 2010-2019) were gathered and pre-processed NLP analysis. A workflow that have been used for the NLP analysis is shown in the Figure 1 below.

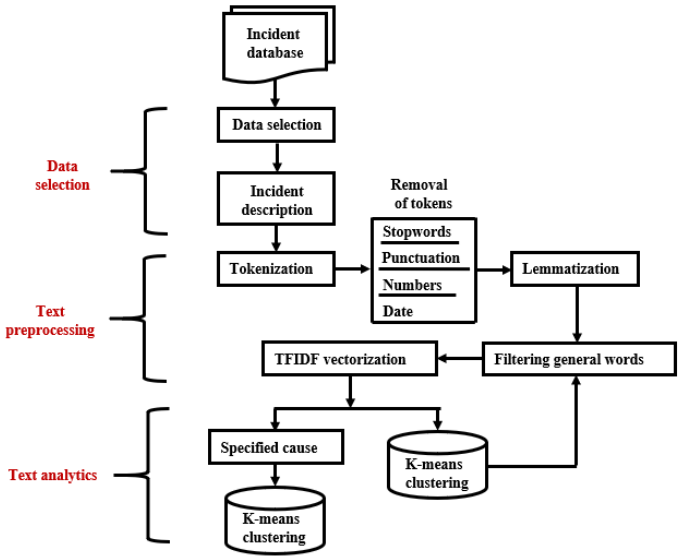

Figure 1. Workflow of natural language processing of incident reports

Development of the artificial neural network (ANN) model: An ANN model is being developed for corrosion failure. The corrosion model has been targeted because there are a lot of corrosion model to compare and validate. Moreover, the number of input data are much more than other failure causes. The ANN model will be expanded upon validation of the corrosion model.

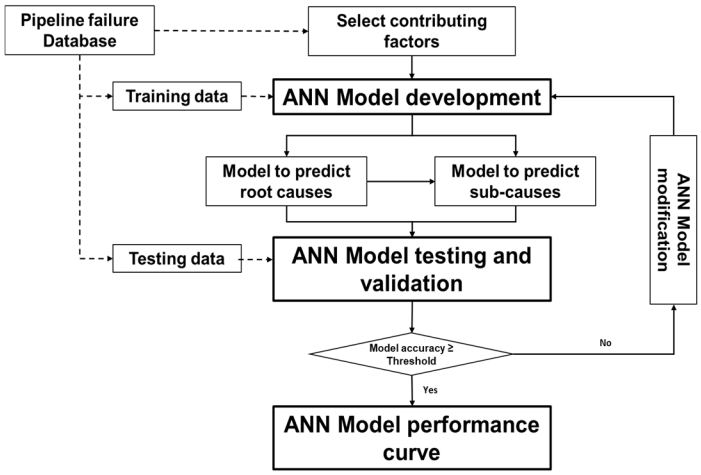

Figure 2. ANN modeling flow

3. Results and Discussions

Literature review: A brief summary of literature review on pipeline incidents is shown in the table below. The table shows the causal factors and background factors that previous studies considered. It is interesting to observe that none of the articles have considered managerial or organizational factors probably because such data are absent largely in the PHMSA or similar

datasets.

Table 1. A brief summary of articles on causal analysis of pipeline incidents

Author (Year)	Causal factors	Background factors	Data source
(Andersen & Misund, 1983)	Outside force/ third party damage, corrosion, mechanical failure, material and construction defects	Pipeline age, location, diameter, commodity transported	CONCAWE, US DOT
(Papadakis, 1999)	Corrosion, external interference, construction/ material defect, other	Pipeline diameter, commodity transported, location	CONCAWE, EGIG, US DOT, VNIIGAS (Soviet Union)
(Bersani, Citro, Gagliardi, Sacile, & Tomasoni, 2010)	Corrosion, mechanical, third-party	*Hydrological, anthropogenic, technical factors	CONCAWE, US DOT
(Han & Weng, 2011)	External interference, corrosion, material defect, operation error, ground movement	Flow rate, pressure, wall thickness, pipeline diameter, service life, depth of cover	US DOT GTG
(Cunha, 2012)	Corrosion, material construction, natural causes, third-party action, others-unknown	Commodity transported, coating type, wall thickness, nominal diameter, population density, depth of cover	EGIG, CONCAWE, UKOPA, US DOT, Trans Petro, NEB
(Wang & Duncan, 2014)	Corrosion, outside force, construction/ material defects	Pipeline age, location	US DOT GTG
(Siler-Evans, Hanson, Sunday, Leonard, & Tumminello, 2014)	Weather/ natural disaster, outside forces, operator error, material failure, corrosion, other		US DOT
(Lam, Zhou, & Piping, 2016)	Corrosion, material failure, excavation damage, other outside forces, natural forces	Location, pipeline material, pipeline age, diameter, corrosion prevention measure	US DOT GTG
(Ramírez-Camacho et al., 2017)	Third-party activity, corrosion, mechanical failure, operational/human error, natural hazards, equipment failure	Pipeline material, population density	MHIDAS
(Bubbico, 2018)	Corrosion failure, equipment failure, excavation failure, incorrect operation, material failure of pipe or weld, natural force damage, other outside force damage, other incident cause	Commodity transported, pipe material, location, corrosion protection system	US DOT

Pipeline failure data analysis: The failure data were compared calculated for all causal factors and compared with that of other failure datasets.

Table 2. Number of pipeline incidents and their percentage distribution for different causal factors for PHMSA HL, PHMSA GTG, NEB, and EGIG datasets are presented. Number in the parenthesis indicates the total number of incident and percentage distribution is shown above that. All failure rates are converted to number of failures per 1,000 km-year.

Data source	US PHMSA HL (2010 – 2019)	US PHMSA GTG (2010 – 2019)	Canada NEB (2008 – 2019)	Europe EGIG (2007 – 2016)

Causal factors	% (#) of incidents	Failure rate /1000 km-year	% (#) of incidents	Failure rate /1000 km-year	% (#) of incidents	Failure rate /1000 km-year	% of incidents	Failure rate /1000 km-year
Corrosion	20.1 (727)	0.227	19.1 (228)	0.048	11.0 (139)	0.163	25.0	0.037
External interference	5.8 (208)	0.065	18.5 (214)	0.045	17.1 (216)	0.253	28.4	0.043
Incorrect operation	14.1 (511)	0.159	5.6 (65)	0.014	10.8 (137)	0.160	3.9	0.006
Equipment failure	45.2 (1635)	0.509	31.4 (363)	0.077	20.5 (259)	0.303	17.8	0.027
Material failure	7.2 (260)	0.081	11.3 (131)	0.028	10.8 (137)	0.160		
Natural force damage	4.5 (161)	0.050	7.9 (92)	0.019	4.7 (60)	0.070	14.9	0.022
Others	3.2 (114)	0.036	5.5 (64)	0.013	3.6 (46)	0.054	10.1	-
Combination factors	-	-	-	-	21.4 (270)	0.316	-	-

To understand the relationship with the background factors such as pipe diameter, it association with corrosion failure has been investigated.

Table 3 Association between pipe diameter and corrosion

Corrosion (Total incident: 752; failure rate: 0.227 failures/1000 km-year)					
Pipe diameter	Mileage	# of incidents	% of incidents	Failure rate	% deviation from average
0 – 6 in	34160	63	8.4	0.122	46.3
6 – 12 in	104641	242	32.2	0.153	32.6
12 – 18 in	29450	85	11.3	0.191	15.9
18 – 24 in	24625	60	8.0	0.161	29.1
24 – in	17930	25	3.3	0.092	59.5
Unknown	218949	277	36.8	0.084	63.0
Total	218949	752		0.227	

Development of natural language processing (NLP) model: After preprocessing K-mean clustering algorithm has been employed on incident description that involved corrosion failure (752 incident counts in HL 2010-2019). The data shows clear division of three categories. Careful observation shows there are one category of external corrosion and two categories of internal corrosions.

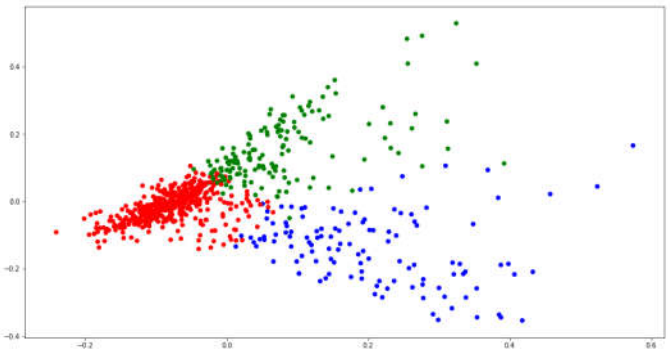

Figure 3. Two-dimensional visualization of clustering results for *corrosion failure*

Publications: Two articles have been accepted for Global Congress on Process Safety presentations and another two articles submitted to other conferences.

- Guanyang Liu, Mason Boyd, and Noor Quddus, Extracting Causal Relations from Incident Reports: A Natural Language Processing and Topic Modeling, (accepted for presentation and poster) 2020 Spring Meeting & 16th Global Congress on Process Safety
- Pallavi Kumari and Noor Quddus, Causation Analysis of Pipeline Incidents Using Artificial Neural Network, (accepted for presentation) 2020 Spring Meeting & 16th Global Congress on Process Safety
- Guanyang Liu, Mason Boyd, and Noor Quddus, Analysis of Pipeline Incident Data and Investigation Reports Using Natural Language Processing (NLP), Abstract accepted, paper submitted to Hazards30, 18-20 May, Manchester, UK
- Pallavi Kumari, Guanyang Liu, Mason Boyd, Syeda Zohra Halim, and Noor Quddus, Causation Analysis of Pipeline Incidents Using Artificial Intelligence, Submitted to International Pipeline Conference (IPC2020), Sep 28- Oct 2, 2020 Calgary, Canada

4. Future work

Future work will be continued on the following

- Literature review will be continued for NLP, ANN and other machine learning applications in pipeline incident data.
- Data analysis will be continued to understand the data that are useful for the ANN analysis.
- The developed NLP model will be applied to identify new root causes or contributing factors and refinement will be conducted.
- The developed NLP model will be applied to incident investigation reports available at PHSMA and NTSB.
- Other NLP techniques will be tested for best extraction of the root causes using the dataset.
- The development of ANN model will be applied to corrosion failure dataset.
- The development of ANN model will be applied to the complete selected failure dataset.